The impact of big data in predictive analytics towards technological development in cloud computing

Krishna Kumar Mohbey* and Sunil Kumar

Department of Computer Science,
Central University of Rajasthan,
Ajmer, 305817, India
Email: kmohbey@gmail.com
Email: 2sunil.cs@gmail.com
*Corresponding author

Abstract: Presently, we are living in a world of data. More than 2.5 quintillion bytes of data are generated everyday. The volume of data from different sources and in various forms can be identified as big data. With the collection of an enormous amount of data, various kinds of predictions would help make decisions. Making intelligent decisions in different situations using massive datasets is known as predictive analysis. Cloud computing aims to be an important way to handle big data. Working on big data in the cloud, however, poses its challenge. Predictive analytics is applied to generate different kinds of patterns that make optimised decisions. This paper introduces big data predictive analytics and its importance. It also gives details of applications and challenges in the present and future scenarios of cloud computing. Besides, we have also included various technologies and frameworks to store, manage, and process big data in cloud platforms.

Keywords: big data; cloud computing; predictive analytics; machine learning approaches; statistical approaches.

Reference to this paper should be made as follows: Mohbey, K.K. and Kumar, S. (2022) 'The impact of big data in predictive analytics towards technological development in cloud computing', *Int. J. Engineering Systems Modelling and Simulation*, Vol. 13, No. 1, pp.61–75.

Biographical notes: Krishna Kumar Mohbey is an Assistant Professor of Computer Science at the Central University of Rajasthan, India. He received his Bachelor's in Computer Application from MCRPV Bhopal in 2006, Master's in Computer Application from Rajiv Gandhi Technological University Bhopal in 2009, and PhD from the Department of Mathematics and Computer Applications from National Institute of Technology Bhopal, India in 2015. His areas of interest are machine learning, data mining, mobile web services, big data analysis, and user behaviour analysis. He has authored three books on different subjects and published more than 25 research articles in reputed journals and conferences.

Sunil Kumar is a scholar at the Central University of Rajasthan, India. He holds a Bachelor's in Computer Application from SMUDE in 2014 and Master's in Computer Application from Patna University in 2017. His research areas include big data analysis, data mining, and machine learning. He has more than five research publications published in prestigious journals and conferences.

1 Introduction

Today, data is available everywhere and generating from various sources. The growth of data has been developing exponentially since the last few years due to the rise of smartphones, the internet of things, and internet technologies. Presently, the use of electronic devices is becoming a necessity for everyone. People are generating data every second in different forms. Another essential source of data is social media applications such as Facebook, Twitter, WhatsApp, etc. These apps generate a massive amount of data every second in various formats that is structured, unstructured, and semi-structured. Due to the volume, velocity, and variety of these vast data, big data is becoming an essential topic in research. There are various challenges to store, manage, compute, and analyse this big data.

In the present day, various companies and organisations are working towards big data storage and computations in productive ways. Multiple technologies are already being developed to handle such kind of important big data. However, they are not able to manage the data efficiently. It is needed to develop effective and practical approaches to address such kind of important big data. The development of tools and frameworks is also important for business

companies. Based on these strategies, decision-making, data analytics, and prediction have become effective.

Big data requires data manipulation of petabytes and maybe exabytes (EXs) and zettabytes (ZBs) of data. The flexible environment of the cloud enables data-intensive applications to be implemented that power business analytics. The cloud also simplifies networking and collaboration within an enterprise, providing access to apply analytics to more workers and streamlining data sharing. Big data analytics, big data management, and big data privacy and security are just a few of the issues that need to be addressed for improved service quality. Blockchain has a lot of potential to develop big data services and applications because of its decentralisation and security (Deepa et al., 2020).

Cloud computing has revolutionised the IT industry by allowing companies to pay only for the tools and services they use, allowing them to be more flexible in their IT consumption. Clouds are often promoted for their ability to provide services on a pay-as-you-go basis, increased availability and elasticity, and cost reduction (Reddy et al., 2014).

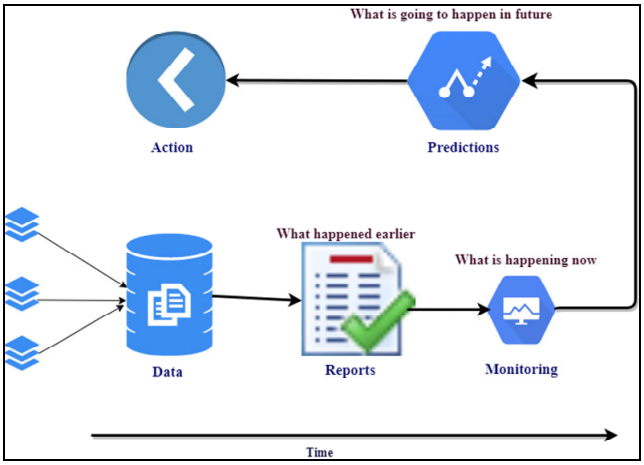
Predictive analytics, on the other hand, is also a critical research topic these days. It is essential for businesses, governments, and organisations. Predictive analytics are also used to make recommendations for specific events. For example, suppose we have historical data of a group of peoples of their daily purchase activities. In that case, the recommender system may predict what kind of new items will be purchased by these people in the future. Accordingly, business companies can provide information and promotional offers to these groups to increase their sale and gain profits.

Predictive analytics needs continuous data related to a particular topic of interest and technologies to make a recommendation. Here, big data tools, techniques, and different frameworks play a vital role in the present day. Based on these technologies, prediction can be performed on a vast amount of big data. Figure 1 shows the process of predictive analytics beginning from data generation to perform actions. It provides the answer to the future, such as 'what will happen in the future', based on the current data availability. According to predictions, future action plans can be made by companies or even individuals. The present paper presents big data, predictive analytics, and their relation to decision-making or recommendations. It also describes the modern tools and technologies related to handling big data to predictive analytics.

This survey's main goal is to demonstrate the impact of big data in predictive analytics, as well as cloud computing and other emerging technologies.

The present paper is organised as follows: Section 2 provides details of big data-related tools, technologies, and various applications. Section 3 includes the concept of predictive analytics. It also includes multiple approaches and frameworks toward predictions. Section 4 discusses the relationship between big data and the predictive analytics process. Section 5 discusses opportunities and challenges dealing with big data and predictive analytics. The last section (Section 6) will present conclusions followed by future research opportunities in this field.

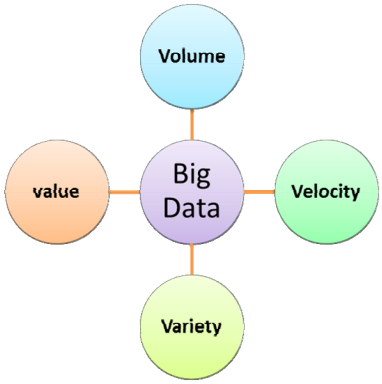
Figure 1 Predictive analytics process (see online version for colours)



2 Big data

Big data refers to the enormous volume of data, structured, unstructured, or semi-structured. In the present scenario, social media and digital devices are the primary sources of big data generation. Almost every individual using smartphones, laptops, and different equipment types generates data in various forms. It may be in the form of a report, document, text, message, tweet, image, audio, video, and so on. These high volumes of generated data cannot be stored, computed, and processed by traditional approaches. Therefore, technological development and enhancement approaches are the necessity of the current time. Decision-making can be done through a systematic process and efficient storage of these big data. While working on big data and related technologies, the first task is to understand big data characteristics, including volume, velocity, variety, and value (Wu et al., 2013).

Figure 2 Four vs. of big data (see online version for colours)



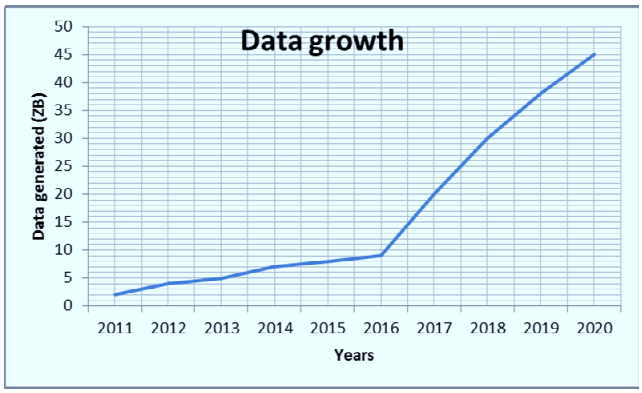
2.1 Characteristics of big data

In recent years, the world we live in is filled with data everywhere; data has grown at scales in diverse areas that we have not encountered on a vast scale. The concept of big data is explicitly used to characterise the exponential growth of structured and unstructured data with the rapid rise in global data. The following four, V's, are the key features of big data (Mohbey, 2019). These features are also shown in Figure 2.

2.1.1 Volume

Volume refers to the massive amount of data that is being generated daily. The primary data generation sources are social media platforms, business processes, machines, human interactions, and networking devices (Oracle, 2017). It is observed that 90% of today's data has been generated in just the last two years. The volume of data can be measured in different units such as byte, KB, MB, GB, TB, PB, EX, ZB, or yottabyte (YB). Figure 3 shows the data growth rate.

Figure 3 Data generation growth rate (see online version for colours)



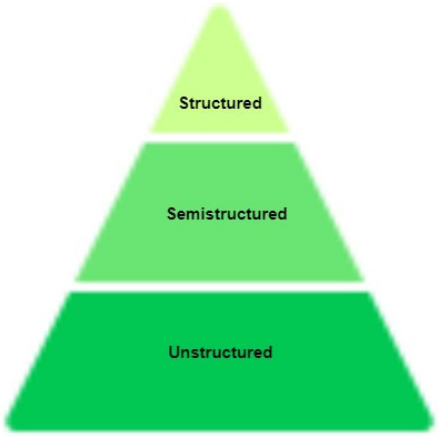
2.1.2 Velocity

Velocity in big data is referred to the speed at which information is being generated in real-time. It is also known as the data in motion. For example, it is the streaming of sensor data, stores in data warehouses, or systems. Real-time data processing requires high data velocity to make immediate decisions, while low data velocity generates and transfers data slowly. Real-time generated data can be used by researchers or business companies to make valuable decisions. It is also used for strategic planning (Mohbey, 2019; Oussous et al., 2018).

2.1.3 Variety

One of the main reasons for the rapid growth of data volume is that data comes from different sources in different formats. Variety refers to the many formats of data that are generated. The generated data may be structured, unstructured, or in the format semi-structured. Earlier data were primarily collected in a structured form, but due to IoT, mobile, and the internet, most of the information is unstructured. Today has received data like emails, files, text, audio, video, pdf, etc. Even publicly, data such as online, weather, finance also available in various formats (Oussous et al., 2018). Figure 4 shows the different variety of generated data.

Figure 4 Different varieties of generated data (see online version for colours)



2.1.4 Value

The ultimate goal of any business is to gain value from big data. It refers to the transformation of raw data into meaningful data. It deals with a mechanism to bring out the correct meaning of the collected data. After collecting data, it is needed to perform mining, then analyse and produce meaningful results. That will be used in business for various decision-making approaches (Mohbey, 2019).

2.2 Traditional data vs. big data

Traditional data is available in a structured form that contains categorical and numerical data. It can be easily stored in RDBMS structured that may be in row and column formats. Its size is enough to manage by systems and computing platforms. It is collected from some specific sources, and all data present in the same form. It can be store, organised, and processed efficiently, while big data is totally different and extremely large. It is also available in various or mixture forms, such as structured, semi-structured, or unstructured (Weber et al., 2014). It cannot be stored and processed on standalone systems. It required a high-performance computing environment or cloud-based architecture to store, manage, and process. It can be in ZB or more. Big data has various qualities, such as highly scalable, flexible, real-time streaming, and quick decision processing (Bharathi and Mandal, 2015; Hazen et al., 2014). Big data is generally produced in semi-structured form, i.e., weblog, clickstream, emails, audio, text, etc. As data availability explodes up, enterprises will need reliable, robust, and automated analytical techniques. Big data analytics offers insight that is useful to extracts hidden patterns from unstructured data. For big data

analysis, technologies such as Hadoop, NoSQL, and Map Reduce are essential (Nema et al., 2018).

Various companies earlier using ERP systems to manage and stored data in RDBMS, where quantity and data format fixed, but big data is different from traditional data. Another essential difference is data analytics. In traditional analytics, data analytics can apply to get results, while big data analytics can provide real-time feedback.

2.3 Big data types

Big data is classified widely into three main types, viz. structured, unstructured, and semi-structured data used for analytics.

2.3.1 Structured data

It refers to the data which is stored in relational databases. It can be easily managed in rows and columns form. Its format is fixed, and processing, storing, retrieval is also more comfortable. This data constitutes about 20% of today's total data and accessible through various database management systems. Machine-generated data and human-generated data are the primary sources of structured data. Table 1 shows an example of structured data.

Table 1 An example of structured data

EmpId	*Name*	*Post*
101	David	Manager
102	John	Clerk
103	Smith	Manager
104	Michel	Peon
105	Joy	Clerk

2.3.2 Unstructured data

Another format of today's data is unstructured; it has no exact format in storage. In today's data, about 80% of the data are available in an unstructured format. Unstructured data may be in different forms, such as image, text, video, documents, etc. The primary sources of unstructured data are smartphones, social networking activities, internet searches, machines, and sensors. This type of data is complicated to process and analyse. Unstructured data cannot be stored using traditional relational databases, and even traditional approaches are not sufficient to process this kind of data.

2.3.3 Semi-structured data

It is a form of structured data that does not conform with the formal structure of data models associated with relational databases or another form of data tables. It may be in the form of structured or unstructured. It includes the data that is not of the traditional database format as structured data but contains some properties which make it easier to process and analysed. For example, NoSQL documents are considered semi-structured since they include keywords that can be used to treat the text quickly. Figure 5 shows semi-structured data.

Figure 5 An example of semi-structured data (see online version for colours)

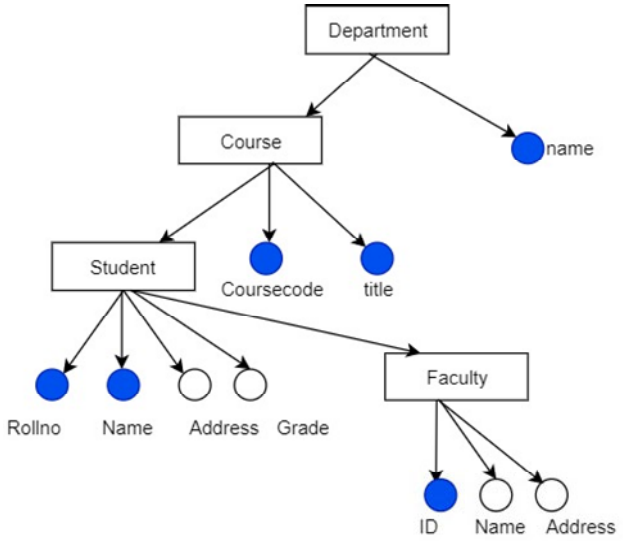
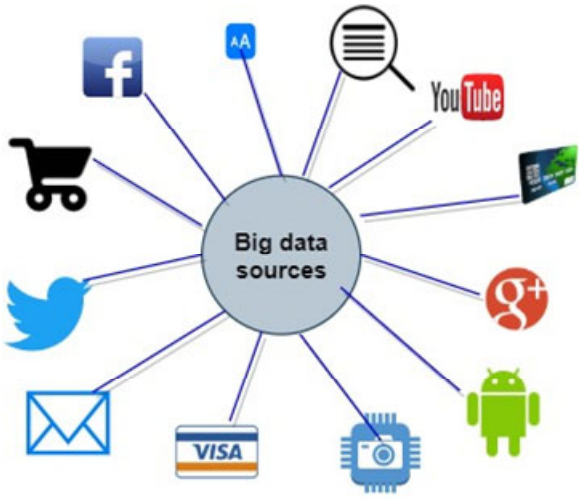


Figure 6 Various sources of big data (see online version for colours)



2.4 Sources of big data

There are many causes to collect big data as it comes in different formats, including social data, machine data, and transactional data. Social media data provides remarkable insights to companies on consumer behaviour and sentiment that can be integrated for data analysis. For example, 230 million tweets are posted on Twitter per day, 2.7 billion Likes and comments are added to Facebook every day, and 60 hours of video uploaded to YouTube every minute (https://onlineitguru.com/tutorial/sources-big-data). Big data is typically generated from one of three primary sources: internet/social networks, traditional business systems, and increasingly from IoT. The data generated from these sources can be structured, unstructured, or semi-structured,

or any combination of these varieties (Guerra et al., 2019). Few sources of big data have shown in Figure 6.

3 Big data technologies

3.1 Distributed processing

Since the digital age advent, data sizes and computation power have increased significantly, but data has risen far beyond processing speeds. The storing and analysis of the conventional approach's data is typically centralised on a single computing device using a particular algorithm. However, for a big data problem, these techniques are insufficient. The imbalance between the exponential pace of data growth and developments in computation and memory processing rates demands distributed processing to cope with huge sets of data (Boja et al., 2012). The purpose of distributed computation is to perform processing based on distributed systems, including data, machines, and techniques. Within a single computer, each algorithm manages its associated set of data leading to local results. These local results are then aggregated to create the final results. Google's MapReduce is the world's first distributed computing cloud-based system. Hadoop is the most used open-source distributed framework.

Parallel computing can be used to provide approaches to handle large-scale dataset processing. Parallel computing is primarily based on splitting the problem or data into smaller parts, which are performed separately by a single processor. The computation is done simultaneously and in parallel. Finally, these outcomes are merged into a single component to generate the final result.

3.2 High-speed local networking

As computing resources, modern parallel computing systems often utilise commodity processors, multicore, multithreaded, or GPUs-based computers (Parhami, 2019). It allows quick improvements to be made to parallel operations with the processor's speed and energy consumption. Therefore, the disparity between different parallel processing systems is primarily dictated by the communication networks used to exchange data between computing nodes. The communication network for computing services, including Ethernet, may also be a convenience. Nonetheless, a custom network or at least an individual commodity switch configuration that meets communications specifications always pays off. Several architectures of network interconnections were developed and used for multiple systems scales over the years. The same electronic chip links computing nodes through an on-chip network, which, given the limits on area and power, is often more restricted in nature. Interlinking computing devices inside an extensive network poses problems in cabling and packaging. Available topologies are inefficient because they cannot be carried out within the physical constraints of partitions, storage, and long cables' signal latency. Interlinking servers in a data centre have fewer topology limitations but dominate concerning size, energy consumption, reliability, or serviceability factors.

3.3 Distributed and parallel data storage

A distributed database is a set of many logically interconnected databases spread over a network of computers (Boja et al., 2012). A distributed database management system is the framework that enables distributed database management and accountability. A parallel database mounted on a computer with multiprocessors. The parallel database applies the horizontal partitioning principle by dividing sections of a big table in a parallel manner over multiple nodes. This needs for partitioned SQL operator's execution. Some fundamental functions, such as a basic SELECT, can be performed on all nodes independently. A multi-operator pipeline carries out more complex operations. Different parallel multiprocessor architectures, such as shared disks, shared memory, or share nothing, are different strategies for parallel database implementation, each with its benefits and disadvantages (Basha and Rajput, 2019). The exchanging nothing method distributes data across separate nodes and has been used for various commercial systems' extensibility and functionality. Based on the definitions above, parallel database systems can be inferred that improve data handling through the concurrent loading, indexing, and querying of data. Existing commercial parallel databases, Teradata, Aster Data and Netezza, Vertica, Greenplum, Oracle Exadata, and IBM DB2, are well established.

3.4 Massive parallel processing

It takes time to process massive datasets from multiple sources because enormous storage and processing capacity is required. The computation and analysis of extensive data with a conventional serial strategy are challenging. The partition of data across different processing units and calculation offers a linear increase in processing power. The massively parallel processing architecture reduces deployment effort because hardware and software are pre-installed and reviewed before they are acquired in data centres. It also eliminates the management effort in the form of a single supplier from the box solution. The data storage devices provide high durability by using data redundancy for each disk by built-in failover space. Ideally, any data processing unit handles the same volume of data at any given time. The data should be uniformly distributed across every processing unit to accomplish this. By doing the same work in every processing unit, all processing units complete their task simultaneously, reducing waiting times. The fact that all data is linked to the same processing unit also significantly affects the results. It reduces the time required for data transmission between the processing units. The distribution of data through parallel database nodes affects overall performance. Although the number of nodes gives the power of the parallel DBMS, this can be an annoyance. For simple queries, the actual time necessary to start the parallel process may be substantially shorter. Nodes may

also transform into hot spots or bottlenecks when the whole network is delayed (Demchenko et al., 2014).

3.5 In-memory database

The big data revolution has driven much work to develop tools for ultra-low-latency and data processing. Given the high access latency to storage devices, current disk-based systems can no longer provide a prompt response. A memory system/database, which holds the random RAM data, is required to meet the strict real-time requirements for analysing mass quantities of data and service requests in milliseconds (Zhang et al., 2015). Primary memory data can be retrieved quicker than disk or flash storage. A variety of in-memory database systems like Microsoft's SQL Server Hekaton, Oracle's TimesTen, MIT's Silo, SAP HANA, and VoltDB eliminate overhead with a focus on ensuring the maximum memory workload efficiency (Graefe et al., 2014).

4 Predictive analytics

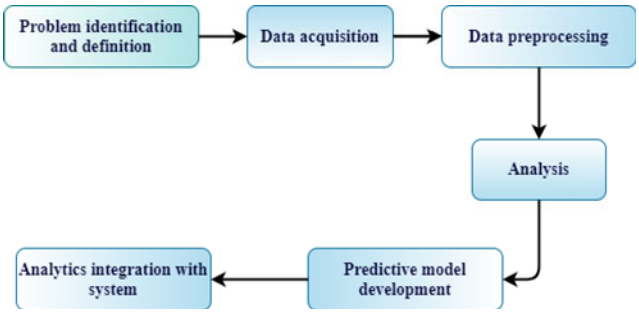
Predictive analytics refers to the making of predictions about future events based on historical data and analytics approaches. It is a branch of data analytics. It uses various tools and techniques of data mining, statistics, modelling, AI, and machine learning. Its demand is rising in recent days due to its capability of predicting future observations and recommendations. Business companies, government organisations, and individuals can make an appropriate action plan if predictive information is available. For example, if a business company knows about the demand for a product in the future, they can manufacture that product in sufficient quantity to supply and gain profits. It is imperative to have appropriate historical and current observed data to make any prediction. Based on these data, various predictive modelling and analytical techniques can apply together for future predictions. Predictive analytics are also helpful in detecting risks and future opportunities for organisations. It provides benefits to the business in competitive environments. Meaningful patterns are generated by predictive analytics that can use to make recommendations and decisions. Almost every company uses predictive analytics in some ways. According to the need, various suggestions and predictions can be made. Finance, banking, healthcare, automobiles, education, aerospace, retailing, hospitality, and manufacturing are critical areas where predictive analytics plays essential roles. Figure 7 shows predictive analytics steps that start from problem identification to model development and integrations towards recommendations.

The following steps used in predictive analytics:

4.1 Problem definition

This step defines the actual goals, outcomes, objectives, scopes, and effort of the process.

Figure 7 Predictive analytics steps (see online version for colours)



4.2 Data acquisition

Historical and current data is the backbone of predictive analytics. In this step, we collect complete data from different sources and combine them. The collected data provides a comprehensive view of the customer or business.

4.3 Data preprocessing

While data collected from multiple sources or devices, there may be some fields or information that are incomplete or have missing values. To process any data, it has to be complete, so that preprocessing is essential. Data cleaning, filling missing values, removing duplicate values, and other tasks are performed in this step.

4.4 Data analysis

It is the process of inspecting, cleansing, transforming, and modelling data with the proposed objectives to make predictions.

4.5 Develop predictive models

It refers to the actual model development towards future event predictions. It is possible to have multiple models out of which the best model can be selected.

4.6 Integrate analytics with system

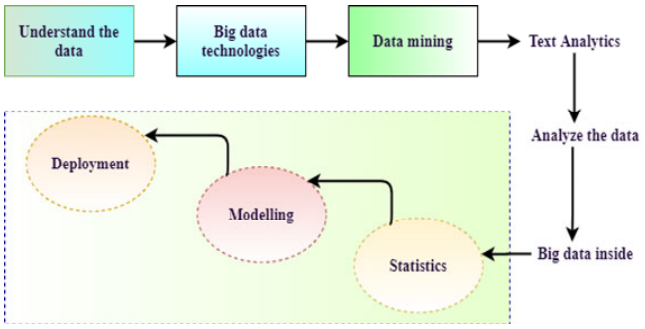
In this step, the proposed models' results deploy to everyday decision-making processes to achieve defined objectives. It is the actual result, report, and output of the predictive analytics system.

5 Big data predictive analytics

The predictive analytics process has been discussed in the previous section for traditional data processing and future observations. However, when we have a vast amount of big data in multiple formats and different velocities, traditional approaches are not adequate for predictive analytics. Therefore, we need efficient frameworks to make predictions using big data and its technologies (Basha and Rajput, 2018; Deepa et al., 2020).

Big data predictive analytics uses various data mining techniques, statistics, and computing approaches to make appropriate future observations. It is a combination of data science, statistics, and computer science approaches. Figure 8 shows the framework of predictive analytics that works on big data.

Figure 8 Big data and predictive analytics processing (see online version for colours)



The following are big data technologies, tools, frameworks, and models used for predictive analytics.

5.1 Big data acquisition technologies

Data acquisition refers to the gathering, filtering, and cleaning of data before storing in any storage solutions towards analytics. Due to size, variety, and real-time processing, infrastructure is a big challenge for big data analytics. It needs an efficient infrastructure that able to handle high dimensions, high transmission, and volumes. It also requires the feature of distributed environments and supports dynamic data structures. Big data can be acquired using the following technologies.

5.1.1 Kafka

Kafka (Lyko et al., 2016) is a distributed messaging framework for publish-subscribe that primarily facilitates continuous high-throughput messaging. Kafka is intended to coordinate offline and online operations with a parallel loading in Hadoop and the capacity to distribute real-time utilisation across a range of computers. The Kafka single cluster processes all data from various sources. It serves as a bridge between asynchronous transmission and live activity. Kafka may also be used for replicating all data for offline use in another data centre. Kafka can be used to feed Hadoop into offline analytics and to monitor internal organisational analytics. Producer, broker, consumer, and topic are the critical concepts of Kafka's structure. Topics are utilised for message feeding. Producers transmit messages to topics and customers that can link to topics and receive messages from these topics.

5.1.2 Flume

Flume (Lakhe, 2016) is a distributed, robust, and available tool to effectively acquire, organise, and transmit vast log data volumes. Flume offers the ability to absorb and archive data into Hadoop for potential analyses, data from various sources, and varying sizes; the customer has the assurance that their data will be generated based on integrant transactions. Flume has an agent-based design that includes the channel, sink, and source components. Source received data from data generators and sent to channels. Channel can work with several sources and sinks. It is the connection between sinks and sources, obtain the data from the source, and keep it before sinks absorb data. The sink is the final part of the flume agent for collecting and sending channel data to the destination. Flume is generally used in streaming data loads, such as tweets created from Twitter, web-based clicking data, or web-based log files.

5.1.3 NiFi

Apache NiFi (Mătăcuță and Popa, 2018) is built for automating data generation and consuming flow in real-time. It facilitates flow-based programming and requires processors and connectors in the graph processing instead of nodes and boundaries. The consumer connects processors to connectors, and how data is manipulated is described. An essential advantage of Nifi is that it is possible to consume any data for individual data incorporating ingestion techniques. Apache NiFi an efficient and robust data processing and distribution framework that is simple to use. Nifi is a reliable choice with a wide range of functionality and a simple-to-use web interface. It is easy to customise and personalisable.

5.1.4 Sqoop

Sqoop (Aravinth et al., 2015) is a platform that offers data transfers to organised data sources, including RDBMS across Hadoop ecosystems. Sqoop is easily configured to move vast volumes of data between HDFS and held structured data stores. It efficiently extracts data from other platforms. It permits data imports into Hadoop from external data stores and client data stores. By using concurrent data sharing, it guarantees fast efficiency. Sqoop effectively facilitates data processing. It also reduces unsustainable stresses on external providers. Sqoop partitioned data by using mappers.

5.1.5 Gobblin

Gobblin (Qiao et al., 2015) is a generic Hadoop data collection tool and the latest open-source product of LinkedIn. Gobblin is defined by three fundamental principles: generality, expansion, and accessibility. Gobblin facilitates a combination of data sources that can be quickly extended. Gobblin, Hadoop's data ingestion tool. Gobblin has been developed to be standardised and extendable such that it is conveniently managed and tracked with a range of data transfer scenarios. At the moment, it used to take over many different data collection pipelines from multiple data outlets at LinkedIn into Hadoop.

5.2 Big data storage technologies

Storage refers to storing and managing data efficiently that satisfies the need for various applications or services. It is also concern with reading and writing data to or from storage devices. Big data storage technologies provide an efficient way to store and access data with satisfying volume, velocity, and variety properties. Big data can be unstructured or semi-structured. Therefore, RDBMS are not sufficient to manage such kinds of data. HDFS, NoSQL, Mongo DB, Cassandra have mostly used technologies for big data storage and management. The following technologies are used for big data storage.

5.2.1 HDFS

The HDFS (Vora, 2011) is a file system developed to store and process data with high-performance data access; the purpose of HDFS is to allow complexity, stream data access, broad datasets, a basic coherence model, etc. HDFS is equipped to store and transmit vast volumes of data (up to 100 sTB) with concurrent access. The storage of HDFS is spread through a node cluster. A HDFS cluster has two node types, i.e., master (name node) and workers (data node). The name node handles the file system namespace, preserves the file system hierarchy, and documents all directory files and folders. The data node stores and restores blocks in compliance with consumer or name node orders.

5.2.2 Hive

Compared to the Hadoop distributed file system (HDFS), Hive provides an interface that enables organised data to be queried using a SQL query language. Hive (Vora, 2011) operates queries via MapReduce job translation. As a result, except for limited databases, hive queries have a substantial latency. Hive advantages include the SQL interface and versatility to build schemas quickly. The schema can be processed separately from the data and only evaluated at query time. This technique is known as schema-on-read as opposed to SQL databases schema-on-write. Therefore, modifying the scheme is a relatively cheap process.

5.2.3 Hbase

HBase has been developed on top of the HDFS to build a broad and scalable, high capacity framework for working with heterogeneous data like non-textual data forms. Hbase is a distributed, highly scalable, and fault-tolerant NoSQL database. The HBase only stores the position information when the actual data stored in HDFS.

5.2.4 MongoDB

MongoDB (Han et al., 2011) is the most comprehensive and most robust database, accommodating multiple forms; MongoDB embraces the JSON schema in storing specific data types. It provides a universal query language for the plurality of operations, such as querying in single table relational databases. MongoDB embraces JSON data structures to handle complicated data typing schemes. Mass storage access high-speed: MongoDB upload speed is ten times that of MySQL as storage reaches 50 GB. Because of these features, MongoDB recommends MongoDB usage instead of a relational database for several ventures with growing results.

5.2.5 Cassandra

Cassandra (Han et al., 2011) is Facebook's open-source database system. Its schema is quite versatile and does not need the database schema's design in the first instance, so inserting or removing a field is a relatively simple one; it provides range queries, i.e., extended queries for keys, and provides robust scalability. Cassandra is the decentralised storage system composed of several storage nodes, replicates a writing process to particular nodes, and routes the read request to a specific node. The scalability can only be achieved by inserting a node for a Cassandra cluster. Furthermore, Cassandra embraces a rich data structure and a smooth query language.

5.2.6 Neo4j

The high-performance NoSql graph database, Neo4j (Prasad and Agarwal, 2016), is a mature, scalable database. It is created in Java and supports the structure of master slaves. Neo4j is an open- and freely available graphical database. It stores data represented as a graph, a set of nodes, and their associations as edges. Such nodes or edges involve other features, which are key/value pairs.

5.3 Big data tools/framework

5.3.1 MapReduce

Mapreduce is a development tool for massive-scale data processing (Dean and Ghemawat, 2010). Applications define a map feature that executes a key/value pair to produce intermediate key/value results and a reduction feature that fuses all intermediate results with the same intermediate key. Mapreduce was created in 2003 at Google to simplify the implementation of an inverted index for search processing. More than 10,000 different software systems, including high-graphical algorithms, text handling, computer training, and statistical machine translation, have been developed using Google MapReduce. Organisations have most used the Hadoop open-source framework of MapReduce. MapReduce parallelises and carries out the program automatically on a large commodity computer cluster. The runtime system provides information on distributing the input data, program compilation on several machines, managing device failures, and managing the inter-machine interaction needed. MapReduce makes it easy to use the distributed system tools for programmers with little expertise with parallel or distributed systems.

5.3.2 Hadoop

Hadoop offers the robust HDFS (Taylor, 2010) with fault-tolerant, and a Java-based API, which enables parallel computation through the cluster nodes using MapReduce. Hadoop allows users to build and execute jobs as a mapper or reducer with any executable. Hadoop also provides job and task trackers that monitor the deployment of the programs on the cluster nodes. Hadoop intends to position data with the machine node automatically. Hadoop schedules map tasks with data of the same node, or at least the same rack. It is a crucial aspect of the success of Hadoop. In April 2008, a 9/10 node cluster Hadoop system broke a world record with less than 3.5 minutes for sorting a terabyte of data. Task failure can be observed, and programs restarted in other safe nodes. To ensure reliability Hadoop replicated data across multiple nodes. Data flow is implied and implicitly managed, unlike other concurrent processing systems. Hadoop simplifies the development of more effective, distributed programs on a cluster of the commodity.

5.3.3 Spark

Apache Spark is a general-purpose, open-source distributed big data analytics programming platform. Due to in-memory processing, caching, and optimised query running, Spark can perform jobs much better than the previous big data platform such as Apache Hadoop, Strom, mahout (Zaharia et al., 2010). Spark provides Python, Java, Scala, and R programming APIs. Spark provides MLlib machine learning, SQL, graph analysis, and stream processing libraries on the spark core's top. Spark offers two major abstracts for parallel computing: resilient distributed datasets (RDD) and concurrent execution of tasks on these datasets. The RDD is the most basic data layout used in Apache Spark. RDDs are immutable sets of dispersed items. These are resilient as an RDD procedure would create a new RDD without modifying the initial. These are distributed as data is spread through many computers in logical partitions. While RDDs are hard to use, all the Spark data structures are regulated with the most granular efficiency. Spark provides multi-step data pipelines to be built utilising directly acyclic graph (DAG) patterns. At higher levels, spark engines create a driver context that defines the application's high-level processing flow.

5.3.4 Flink

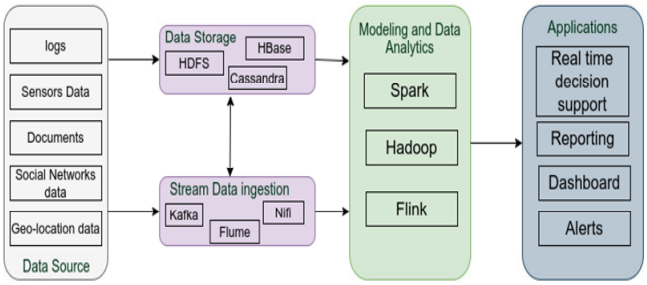
Apache Flink implements the data stream management method as the integrated method for real-time processing, continuous streams, and batch processing. Flink is designed on top of datasets (collections of unique elements specified by an inferred form parameter), job graphs, and parallelisation contracts (Carbone et al., 2015). Job graphs illustrate concurrent data flows that collect and build data streams with arbitrary tasks. Parallelisation contracts are functions that specify properties on the data for their related user-defined purposes, which are used to implement optimisation regulations and parallelise the UDFs execution. Flink's API has two different iteration operators to define iterations:

1 mass iterations, which are conceptually identical to loop unrolling

2 delta iterations in which the set of solutions is altered by steps rather than complete recomputation.

Delta iterations can accelerate dramatically, as the function in each iteration reduces as the amount of repetitions continues. Flink algorithms can measure both early and preliminary and delayed and precise resulting in a highly robust windowing process. Flink can provide programmers with a considerable degree of freedom to determine how events are handled by various notions of time (ingestion-time, event-time, processing time). Flink has a specific API that utilises particular data structures and algorithms to manage static datasets for batch variants of operators, such as joining or grouping and utilising timelines. The effect is that Flink is a complete, powerful batch processor, including the libraries for graph analysis and machine learning, on top of a streaming runtime. Figure 9 shows how big data processing tools can be used in the framework of predictive analytics.

Figure 9 Big data framework for predictive analytics (see online version for colours)



6 Predictive modelling techniques for big data

Predictive modelling is a combination of statistics, mathematics, and computer sciences. It also required domain experts who can handle business problems. It may include experts from engineering, medical science, economics, and so on. Predictive modelling techniques can be classified in several ways. The selection of modelling techniques depends on the number of parameters, such as the field of problem, dataset, and what kind of outcome we expect. The following are the broader categories of predictive modelling techniques that use big data as an input parameter.

6.1 Supervised learning models

In supervised models, the modeller uses a dataset where the values of response variables are known. This kind of model uses historical data to train their model, and prediction can be made for a new set of data. Various applications these

days use such types of modelling techniques for forecasts. For example, if we want to predict a transaction is fraudulent or not in online banking. Supervised learning can also be classified as classification and regression (Iwendi et al., 2019).

6.1.1 Classification models

Classification models are used to predict the categorical outcome from known categories. The expected result may be a binary response, i.e., 0, 1, or multi types. Naïve Bayes, support vector machine (SVM), decision tree, random forest, and neural networks are examples of classification models (Kolisetty and Rajput, 2020).

6.1.2 Regression models

Regression models are used to predict a continuous outcome from available features. The continuous values can be numeric or quantitative that have to be anticipated and not from existing categories. Linear regression is an example of a regression model that is used to predict the continuous outcome efficiently.

6.2 Unsupervised learning models

Unsupervised learning modelling does not attempt to predict a specific result as supervised models predict but allows models to work independently to discover information. It mainly deals with unlabelled data. It uses to perform more complex processing tasks. Unsupervised learning is further categorised as clustering models and dimensionality reduction models.

6.2.1 Clustering models

The clustering models combine data into similar categories known as clusters. It is performed based on the same attributes. The objectives of clustering models are to find patterns and provide more information. For example, a retail company may apply clustering models to make different customer clusters based on their purchasing behaviours. K-mean clustering and DBSCAN are examples of clustering models.

6.2.2 Dimensionality reduction models

The process of reducing dimension from feature sets is known as dimensionality reduction (Reddy et al., 2020). It is beneficial when we have a large number of dimensions in the collected data. These models' objective is to provide the highest amount of information from the lowest set of features. Principal component analysis (PCA) and linear discriminant analysis (LDA) are examples of dimensionality reduction models (Reddy et al., 2020).

6.3 Semi-supervised learning models

In this type of learning model, the algorithms are trained upon categorised and unlabelled data. These kinds of modelling techniques are used to enhance supervised models. In recent days, internet content classification, speech analysis, and protein sequence classifications are semi-supervised modelling applications.

6.4 Big data processing using cloud computing

Big data is all about interacting with the large-scale of data, while cloud computing is about infrastructure and services. The fundamental explanation for their tremendous enterprise acceptance is the simplification provided by big data and cloud technologies (Reddy et al., 2014). For instance, Amazon's 'Elastic Map Reduce' shows the power of cloud elastic computes for big data processing. Cloud computing provides consumers with services based on a pay-as-you-go model (Mohbey, 2017). Cloud providers offer three primary services, and these services are listed below:

6.4.1 Infrastructure as a service (IAAS)

Here the service provider provides the whole infrastructure along with the activities associated with maintenance. It provides computing resources, such as storage or computation.

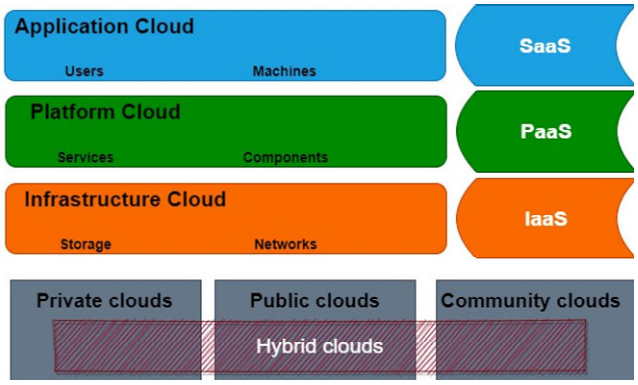
6.4.2 Platform as a service (PAAS)

The cloud provider provides tools for this service, such as object storage, runtime, queuing, databases, etc. Nevertheless, the responsibility for configuring and performing relevant activities rests on the user (Abou El-Seoud et al., 2017; Cole et al., 2019).

6.4.3 Software as a service (SAAS)

This service is the most convenient one that provides all the requisite settings and provides IaaS infrastructure for its infrastructure. It focuses on the actual applications accessed and used by the users (Sriram and Khajeh-Hosseini, 2010).

Figure 10 Big data framework for predictive analytics (see online version for colours)



An overview of the typical deployment and implementation of cloud computing is given in Figure 10. The three models on top of each of the four implementation models can be deployed.

7 Benefits of big data analysis in a cloud environment

Big data processing in the cloud computing environment has many advantages. Some of these are discussed below.

7.1 Improved analysis

Big data analysis has become more effective with advancements in cloud computing technologies, such as better performance and high scalability. Businesses, therefore, tend to conduct big data processing in the cloud computing environment. Besides, the cloud allows incorporating information from various sources and places (Abou El-Seoud et al., 2017).

7.2 Decreasing the cost

Big data and cloud computing technologies provide value to companies by decreasing ownership. Without large-scale big data infrastructure, the cloud environment helps clients to process big data. Therefore for business purposes, both big data and cloud computing technologies push the cost down and add value to the enterprise.

7.3 Basic infrastructure

Big data analysis is a too hectic task as the data arrives in vast quantities of varying speeds and formats that conventional infrastructures are typically unable to keep up with. It is easy to handle workloads as cloud computing offers a flexible infrastructure that we can scale according to requirements at the time (Sriram and Khajeh-Hosseini, 2010).

8 Applications of big data predictive analytics

Big data predictive models may gain insights into many specific dimensions of an organisation, including customers' actions, the manufacturer, and the processor's speed. The following are essential applications of big data predictive analytics.

8.1 Healthcare monitoring

Healthcare and medical data are complex data that are constantly and rapidly increasing, containing abundant and varied knowledge values. Big data predictive analytics has an enormous ability to capture, process, and analyse medical information efficiently. The application of predictive analytics of health care significantly impacts the health care industry, such as using pattern identification algorithms to recognise asthma and COPD (Sanchez-Morillo et al., 2016). An asthma monitoring tool monitors and analyses patients' breathing sounds and offers immediate guidance to support patients who handle asthma and COPD using a mobile phone app. Patients and healthcare professionals can conveniently access their legitimate records regardless of geographic location, time, and cost-effectively via a decision support system such as a cloud-based clinical decision support system (CCDSS) (Oyenuga et al., 2021).

8.2 Predicting customer behaviour

Digital data includes what consumers see, what they read, their behaviour and actions, their opinion, their desires, and expectations to provide a vast volume of data for learning opportunities that can be collected. The big data interest resides in the outcomes of analysis and projections or behaviour derived from the analysing and forecasting results (Surendro, 2019). Customer behaviour can be predicted with predictive analytics when businesses process customer review data and use it for marketing campaigns.

8.3 Energy production

Big data predictive analytics tools are essential components of the smart grid period (Diamantoulakis et al., 2015; Mack, 2014) to maximise growing industry competitors' interest and improve resource usage. Big data predictive modelling, though, is also a long way from being used as a critical factor in implementing smart grid developments. It is also evident that big data predictive analytics energy businesses potentially deliver benefits. There is an ability to create innovative approaches in developing sophisticated forecasting applications models that track plant supply, historical patterns, seasonality, and temperature.

8.4 Current business management

Identify and maintain consumers; introduce different items to existing customers, consumer profiling; improve demand consistency, the initial effect of market divergence in pricing

8.5 Financial services

Developing credit risk systems using big data predictive analytics can improve financial firms to forecast credit danger.

8.6 Transportation

With the development of big data technologies, transportation has become more efficient and effective. Predictive analytics can be used across multiple transportation areas such as route planning, congestion prediction, traffic prediction and management, road monitoring, etc.

8.7 eLearning using cloud computing

Education worldwide, including those in developing countries, is rapidly adopting cloud-based services due to the advantages of cloud computing and the adoption of technology by competitors. The need for educational institutions, in general, and higher education (HE) institutions, in particular, to implement cloud computing applications has increased as a result of the COVID-19 pandemic to engage students in an online mode and conduct research remotely. The adoption of cloud computing in various industries, including HE, has gained traction in developing countries in recent years (Bhardwaj et al., 2021; Njenga et al., 2019).

9 Challenges and opportunities of predictive analytics

9.1 Challenges of predictive analytics

There are some challenges in predictive analytics while using big data as a critical point. The main problem is high dimensionality and the massive size of data that is generating by various services. Organisations and companies need to collect data very fastly because data is making from multiple sources simultaneously. Few other challenges are also included in big data processing, such as noise diversity of data, cybersecurity (Tang et al., 2017), spurious correlation, and measurement errors (Fan et al., 2014). For effective prediction and data analytics, it is essential to deal with data noise, diversity, inefficiency, and correlation. The following are the main challenges related to big data predictive analytics.

9.1.1 High dimensionality

Big data always has high dimensional, and it is challenging to select appropriate features for predictive models. There may be a spurious correlation if the dimensions are chosen wrongly and used in the proposed model. The model may give inaccurate results of the prediction. Another issue is related to the computational cost and algorithm complexity. If high dimensions are combined with extensive data samples, it may need high computational cost and an effective algorithm (Jeble et al., 2016; Iwendi et al., 2018).

9.1.2 Data size

Due to the vast volume of big data, its management is also the biggest challenge for companies and organisations (Patel et al., 2020). Traditional application and services generation limited the size of data that can be easily managed by RDBMS software. In contrast, big data requires specialised frameworks such as Hadoop, NoSQL, Spark, etc. It also needs advanced infrastructure because it has a faster generation, transfer, storage, and processing rate.

9.1.3 Data quality

Completeness of data, accuracy, and availability of data can also relate to the challenge of big data for quality measurement. Big data requires sophisticated approaches to deal with the quality of big data. If data quality is ignored, predictive results will be inaccurate.

9.1.4 Completeness of data

Due to the large dataset and having high dimensions, getting complete data is not possible. Another reason for data incompleteness is heterogeneity in data. Some data may be in a structured format, while some in an unstructured format. Due to hugeness, we may have only subsets of data. It is not enough to make appropriate predictions. Therefore, it needs to have complete data before processing and modelling.

9.1.5 Reliability of data

We collect data from different sources such as smartphones, social networks, text messages, and emails. These data are generally unstructured, and most of the unstructured data is usually unreliable due to data loss. Sometime data also comes from various sensors. If any sensor stops sending data, then there may be data loss; it is not reliable to process for modelling. Missing data may also cause data unreliability. It's because data from the devices or sensors cannot be obtained for a variety of reasons. Multiple data mining processes are required to achieve data reliability (Asif et al., 2013; Garg et al., 2013). This process generates meaningful information from a large volume of data. After this process, modelling can be performed (Le and Pang, 2013).

9.2 Opportunities in predictive analytics and cloud computing

Big data and cloud computing started a new paradigm for researchers to think about in technological development. It generates various opportunities for business companies, organisations, and individuals as well. Multiple companies are entirely dependent on big data, cloud computing, and predictive models. They have started collecting customer information, purchase history, transactions, locations, interests, and reviews to enhance their predictive models. Based on the predictions, they provide various recommendations and offers to motivate and interact with customers. The following are some significant opportunities for future researchers.

9.2.1 Understanding big data and cloud architecture

These days we have superfast machines, massive data, and the latest technologies, but they are not enough to make

appropriate predictive analytics. The first step is to understand and interpret data accurately to make efficient decisions according to data availability. After that, we can import these data to develop predictive models. Various advanced technologies such as Hadoop, Pig, Hive, MapReduce, Spark, and NoSql are available to manage data (Jeble et al., 2016) better in the cloud platforms.

9.2.2 Identification of big data

Identify which data to use is a crucial challenge for model success. Data is coming from different sources in different formats such as transactions, sales, social media, reviews, feedback, etc. It is a more significant opportunity for model developers to select the right data features to develop predictive models. If the wrong element is selected, the model may be failing or give the wrong results.

9.2.3 Data mining

Data mining is a process of identifying hidden patterns from available datasets. It helps to predict future outcomes. Due to heterogeneity and large-scale data, the data mining process needs advanced techniques for pattern identification. There are lots of improvement can be possible in data mining techniques to predict future outcomes.

9.2.4 Business opportunities

In the development of internet technologies, smartphones, and IoT, the business process is completely changed. Most of the businesses are online now. The growth in technology and big data has also started competitions among companies. Therefore, a lot of new opportunities are available in the business. It includes customer satisfaction, timely delivery of the product, cost reduction, and many more. Based on the available data, predictive modelling business companies can provide better services and offers to customers and explore revenue opportunities.

9.2.5 Technological development

Big data and cloud computing both have various opportunities for growth in technologies. Traditional approaches, frameworks, and models are not appropriate to predict future outcomes from the vast size and heterogeneous data. It is an opportunity for companies and researchers to develop advanced methods, frameworks, and models for predictive analytics. Data science, statistics, mathematics, artificial intelligence, and machine learning approaches can be used together to enhance models and techniques. These approaches can produce optimise algorithms to process large datasets with high dimensions (Fan et al., 2014).

9.2.6 Real-time processing

Big data can be used to provide real-time information to those who can derive value from available datasets using cloud computing services. There are several examples where real-time processing plays an essential role. It includes disaster information, weather forecasting, transportations, healthcare, retails, decision making, and so on (Yin et al., 2016; Rai et al., 2015). Various customer services also required instant actions, and it is possible with real-time data availability and processing. For example, based on geolocation mobile phone data, traffic jams can be predicted and provide alternate routes to the user immediately. These kinds of predictions have various opportunities, such as data unavailability, internet issues, weather conditions, road conditions, and so on.

9.2.7 Future directions

The research topics are not limited to big data technical advancements. The key aim is to transform the cloud into a scalable data analytics tool, rather than just a data storage and technology platform. The development of standards and APIs that enable users to easily migrate between solutions, as well as the ability to take advantage of the Cloud infrastructure's elasticity capability. The latter includes expressive languages that allow users to define a problem in simple terms while decomposing a high-level definition into multiple concurrent subtasks and maintaining high performance efficiency even if the problem is complex.

10 Conclusions

Big data and cloud computing is an essential element for future outcomes predictions. The predicted outcome can be used to make appropriate recommendations in various fields. Big data-based predictive analytics is an interdisciplinary field that combines knowledge of data science, statistics, mathematics, and computational sciences with that field's expertise. Predictive analytics uses various modelling techniques and big data technologies such as Hadoop, Spark, and MapReduce. They used to make predictions about any outcome in the future, and recommendations can be made accordingly. The predictive analytics procedure uses historical data and various machine learning approaches, such as classification and regression. In this paper, big data, sources, technologies, and predictive strategies have been discussed. Applications, opportunities, and challenges of predictive analytics are also presented concerning a big data environment. In this paper, various issues of predictive analytics and big data are included. These issues are open challenges for every researcher. Novel approaches and models could be developed for predictive analytics modelling that can be effective and suitable for big data.

References

Abou El-Seoud, S., El-Sofany, H.F., Abdelfattah, M.A.F. and Mohamed, R. (2017) 'Big data and cloud computing: trends and challenges', *International Journal of Interactive Mobile Technologies (iJIM)*, Vol. 11, No. 2, pp.34–52.

Aravinth, S.S., Begam, A.H., Shanmugapriyaa, S., Sowmya, S. and Arun, E. (2015) 'An efficient HADOOP frameworks SQOOP and ambari for big data processing', *Int. J. Innov. Res. Sci. Technol*., Vol. 1, No. 10, pp.252–255.

Asif, M.T. et al. (2013) 'Low-dimensional models for missing data imputation in road networks', *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE.

Basha, S.M. and Rajput, D.S. (2018) 'A supervised aspect level sentiment model to predict overall sentiment on tweeter documents', *International Journal of Metadata, Semantics and Ontologies*, Vol. 13, No. 1, pp.33–41.

Basha, S.M. and Rajput, D.S. (2019) 'A roadmap towards implementing parallel aspect level sentiment analysis', *Multimedia Tools and Applications*, October, Vol. 78, No. 20, pp.29463–29492.

Bharathi, V. and Mandal, T. (2015) 'Prioritising and ranking critical factors for sustainable cloud ERP adoption in SMEs', *Int. J. Autom. Logist*., Vol. 1, No. 3, pp.294–316.

Bhardwaj, A.K. et al. (2021) 'E-learning during COVID-19 outbreak: cloud computing adoption in Indian public universities', *Computers, Materials & Continua*, Vol. 66, No. 3, pp.2471–2492, https://doi.org/10.32604/cmc.2021.014099.

Boja, C., Pocovnicu, A. and Batagan, L. (2012) 'Distributed parallel architecture for' big data'', *Inform. Econ*., Vol. 16, No. 2, p.116.

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S. and Tzoumas, K. (2015) 'Apache flink: stream and batch processing in a single engine', *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng*., Vol. 36, No. 4.

Cole, T. et al. (2019) 'Investigation into cloud computing adoption within the hedge fund industry', *Journal of Cases on Information Technology (JCIT)*, Vol. 21, No. 3, pp.1–25.

Dean, J. and Ghemawat, S. (2010) 'MapReduce: a flexible data processing tool', *Commun. ACM*, Vol. 53, No. 1, pp.72–77.

Deepa, N. et al. (2020) *A Survey on Blockchain for Big Data: Approaches, Opportunities, and Future Directions*, arXiv preprint arXiv:2009.00858.

Demchenko, Y., De Laat, C. and Membrey, P. (2014) 'Defining architecture components of the big data ecosystem', in *2014 International Conference on Collaboration Technologies and Systems (CTS)*, pp.104–112.

Diamantoulakis, P.D., Kapinas, V.M. and Karagiannidis, G.K. (2015) 'Big data analytics for dynamic energy management in smart grids', *Big Data Res*., Vol. 2, No. 3, pp.94–101.

Fan, J., Han, F. and Liu, H. (2014) 'Challenges of big data analysis', *Natl. Sci. Rev*., Vol. 1, No. 2, pp.293–314.

Garg, L. et al. (2013) 'Tensor-based methods for handling missing data in quality-of-life questionnaires', *IEEE Journal of Biomedical and Health Informatics*, Vol. 18, No. 5, pp.1571–1580.

Graefe, G. et al. (2014) 'In-memory performance for big data', *Proceedings of the VLDB Endowment*, Vol. 8, No. 1, pp.37–48, DOI: 10.14778/2735461.2735465.

Guerra, F., Sottovia, P., Paganelli, M. and Vincini, M. (2019) 'Big data integration of heterogeneous data sources: the re-search alps case study', in *2019 IEEE International Congress on Big Data (BigDataCongress)*, pp.106–110.

Han, J., Haihong, E., Le, G. and Du, J. (2011) 'Survey on NoSQL database', in *2011 6th International Conference on Pervasive Computing and Applications*, pp.363–366.

Hazen, B.T., Boone, C.A., Ezell, J.D. and Jones-Farmer, L.A. (2014) 'Data quality for data science, predictive analytics, and big data in supply chain management: an introduction to the problem and suggestions for research and applications', *Int. J. Prod. Econ*., Vol. 154, pp.72–80, DOI: 10.1016/j.ijpe.2014.04.018.

Iwendi, C., Ponnan, S., Munirathinam, R., Srinivasan, K. and Chang, C.Y. (2019) 'An efficient and unique TF/IDF algorithmic model-based data analysis for handling applications with big data streaming', *Electronics*, November, Vol. 8, No. 11, p.1331.

Iwendi, C., Zhang, Z. and Du, X. (2018) 'ACO based key management routing mechanism for WSN security and data collection', *2018 IEEE International Conference on Industrial Technology (ICIT)*, IEEE.

Jeble, S., Kumari, S. and Patil, Y. (2016) 'Role of big data and predictive analytics', *Int. J. Autom. Logist*., Vol. 2, No. 4, pp.307–331.

Kolisetty, V.V. and Rajput, D.S. (2020) 'A review on the significance of machine learning for data analysis in big data', *Jordanian Journal of Computers and Information Technology (JJCIT)*, March, Vol. 6, No. 1, pp.41–57.

Lakhe, B. (2016) 'Implementing SQOOP and flume-based data transfers', in *Practical Hadoop Migration*, pp.189–205, Springer.

Le, C.V. and Pang, C.K. (2013) 'An energy data-driven decision support system for high-performance manufacturing industries', *Int. J. Autom. Logist*., Vol. 1, No. 1, pp.61–79.

Lyko, K., Nitzschke, M. and Ngomo, A-C.N. (2016) 'Big data acquisition', in *New Horizons for a Data-Driven Economy*, pp.39–61, Springer, Cham.

Mack, P. (2014) 'Chapter 35-big data, data mining, and predictive analytics and high performance computing', *Renew. Energy Integr. Acad. Press. Bost*., pp.439–454.

Mătăcuţă, A. and Popa, C. (2018) 'Big data analytics: analysis of features and performance of big data ingestion tools', *Inform. Econ*., Vol. 22, No. 2, pp.25–34.

Mohbey, K.K. (2017) 'The role of big data, cloud computing and IoT to make cities smarter', *International Journal of Society Systems Science*, Vol. 9, No. 1, pp.75–88.

Mohbey, K.K. (2019) 'An efficient framework for smart city using big data technologies and internet of things', in *Progress in Advanced Computing and Intelligent Engineering*, Springer, pp.319–328.

Nema, R., Tandon, J. and Thakral, A. (2018) 'Predictive analytics in big data & intelligent automation', in *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, pp.437–441.

Njenga, K. et al. (2019) 'The cloud computing adoption in higher learning institutions in Kenya: hindering factors and recommendations for the way forward', *Telematics and Informatics*, Vol. 38, pp.225–246, DOI: 10.1016/j.tele.2018.10.007.

Oracle (2017) *An Enterprise Architect's Guide to Big Data* [online] http://www.oracle.com/technetwork/topics/entarch/articles/oea-big-data-guide-1522052.pdf (accessed 15 September 2017).

Oussous, A., Benjelloun, F-Z., Lahcen, A.A. and Belfkih, S. (2018) 'Big data technologies: a survey', *J. King Saud Univ. Inf. Sci*., Vol. 30, No. 4, pp.431–448.

Oyenuga, S.O. et al. (2021) 'Cloud-based clinical decision support system', *Conference Proceedings of ICDLAIR2019*, Springer International Publishing.

Parhami, B. (2019) 'Parallel processing with big data', in Sakr, S. and Zomaya, A.Y. (Eds.): *Encyclopedia of Big Data Technologies*, Springer International Publishing, Cham, pp.1253–1259, DOI: 10.1007/978-3-319-77525-8_165.

Patel, H. et al. (2020) 'A review on classification of imbalanced data for wireless sensor networks', *International Journal of Distributed Sensor Networks*, Vol. 16, No. 4, DOI: 1550147720916404.

Prasad, B.R. and Agarwal, S. (2016) 'Comparative study of big data computing and storage tools', *Int. J. Database Theory Appl*., Vol. 9, No. 1, pp.45–66.

Qiao, L. et al. (2015) 'Gobblin: unifying data ingestion for Hadoop', *Proc. VLDB Endow*., Vol. 8, No. 12, pp.1764–1769.

Rai, S.S., Sharma, V. and Ganguly, K. (2015) 'Logistics complexity in Indian garment supply chain', *Int. J. Autom. Logist*., Vol. 1, No. 4, pp.419–430.

Reddy, G.T. et al. (2014) 'Employing data mining on highly secured private clouds for implementing a security-as-a-service framework', *J. Theor. Appl. Inf. Technol*., Vol. 59, No. 2, pp.317–326.

Reddy, G.T. et al. (2020) 'Analysis of dimensionality reduction techniques on big data', *IEEE Access*, Vol. 8, pp.54776–54788, DOI: 10.1109/ACCESS.2020.2980942.

Sanchez-Morillo, D., Fernandez-Granero, M.A. and Leon-Jimenez, A. (2016) 'Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: a systematic review', *Chron. Respir. Dis*., Vol. 13, No. 3, pp.264–283.

Sriram, I. and Khajeh-Hosseini, A. (2010) *Research Agenda In Cloud Technologies*, 19 January, arXiv preprint arXiv: 1001.3259.

Surendro, K. (2019) 'Predictive analytics for predicting customer behavior', in *2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT)*, pp.230–233.

Tang, M., Alazab, M. and Luo, Y. (2017) 'Big data for cybersecurity: vulnerability disclosure trends and dependencies', *IEEE Transactions on Big Data*, Vol. 5, No. 3, pp.317–329.

Taylor, R.C. (2010) 'An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics', in *BMC Bioinformatics*, Vol. 11, No. S12, p.S1.

Vora, M.N. (2011) 'Hadoop-HBase for large-scale data', in *Proceedings of 2011 International Conference on Computer Science and Network Technology*, Vol. 1, pp.601–605.

Weber, G.M., Mandl, K.D. and Kohane, I.S. (2014) 'Finding the missing link for big biomedical data', *Jama*, Vol. 311, No. 24, pp.2479–2480.

Wu, X., Zhu, X., Wu, G-Q. and Ding, W. (2013) 'Data mining with big data', *IEEE Trans. Knowl. Data Eng*., Vol. 26, No. 1, pp.97–107.

Yin, X.F., Fu, X., Ponnambalam, L. and Goh, R.S.M. (2016) 'A k-means clustering for supply chain risk management with embedded network connectivity', *Int. J. Autom. Logist*., Vol. 2, Nos. 1–2, pp.108–121.

Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S. and Stoica, I. (2010) 'Spark: cluster computing with working sets', *HotCloud*, Vol. 10, No. 10, p.95.

Zhang, H., Chen, G., Ooi, B.C., Tan, K-L. and Zhang, M. (2015) 'In-memory big data management and processing: a survey', *IEEE Trans. Knowl. Data Eng*., Vol. 27, No. 7, pp.1920–1948.